

SUPPLEMENTAL MATERIALS

for

The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease

Table of contents

I. Supplementary Figures

II. Supplementary Tables

III. Supplementary Note

1. Ferret genome assembly
 2. Expanded annotation for differential expression analysis
 3. Influenza virus detection by titer, RT-PCR and immunohistochemical staining
 4. Ferret microarray performance assessment
- Reference

I. Supplementary Figures

Supplementary Figure 1. View within the UCSC genome browser, showing a segment of dog chromosome 9 with the LiftOver of the ferret genome sequence indicated by the track labeled “ferret”

Supplementary Figure 2. Phylogenetic tree based on protein sequences of 789 orthologous groups with representation in at least 31 of the 34 species

Supplementary Figure 3. Gene Ontology enrichment for angular sector representing the top 25% of genes where the ferret sequence is closer to human than the mouse orthologue (see Figure 1a in main text)

Supplementary Figure 4. Protein sequence comparisons of orthologous trios (human, ferret, mouse); y axis: human vs. mouse distance; x axis: human vs. ferret distance

Supplementary Figure 5. Ferret tracheal immunohistochemistry

Supplementary Figure 6. Ferret lung immunohistochemistry

Supplementary Figure 7. Characterization of intergenic transcripts derived from ferret RNA-seq data

Supplementary Figure 8. Comparison of transcript abundances measured by Total RNA-seq vs. mRNA-seq analysis of the same set of ferret lung samples

Supplementary Figure 9. Comparison of microarray technical replicates (ferret microarray version 1)

Supplementary Figure 10. Comparison of microarray technical replicates (ferret microarray version 2)

Supplementary Figure 11. Ferret microarray (version 1) vs. RNA-seq comparison (lung mRNA-seq)

Supplementary Figure 12. Ferret microarray (version 1) vs. RNA-seq comparison (lung Total RNA-seq)

Supplementary Figure 13. Ferret microarray (version 1) vs. RNA-seq comparison (trachea Total RNA-seq)

Supplementary Figure 14. Ferret microarray (version 2) vs. RNA-seq comparison (lung mRNA-seq)

Supplementary Figure 15. Ferret microarray (version 2) vs. RNA-seq comparison (lung Total RNA-seq)

Supplementary Figure 16. Ferret microarray (version 2) vs. RNA-seq comparison (trachea Total RNA-seq)

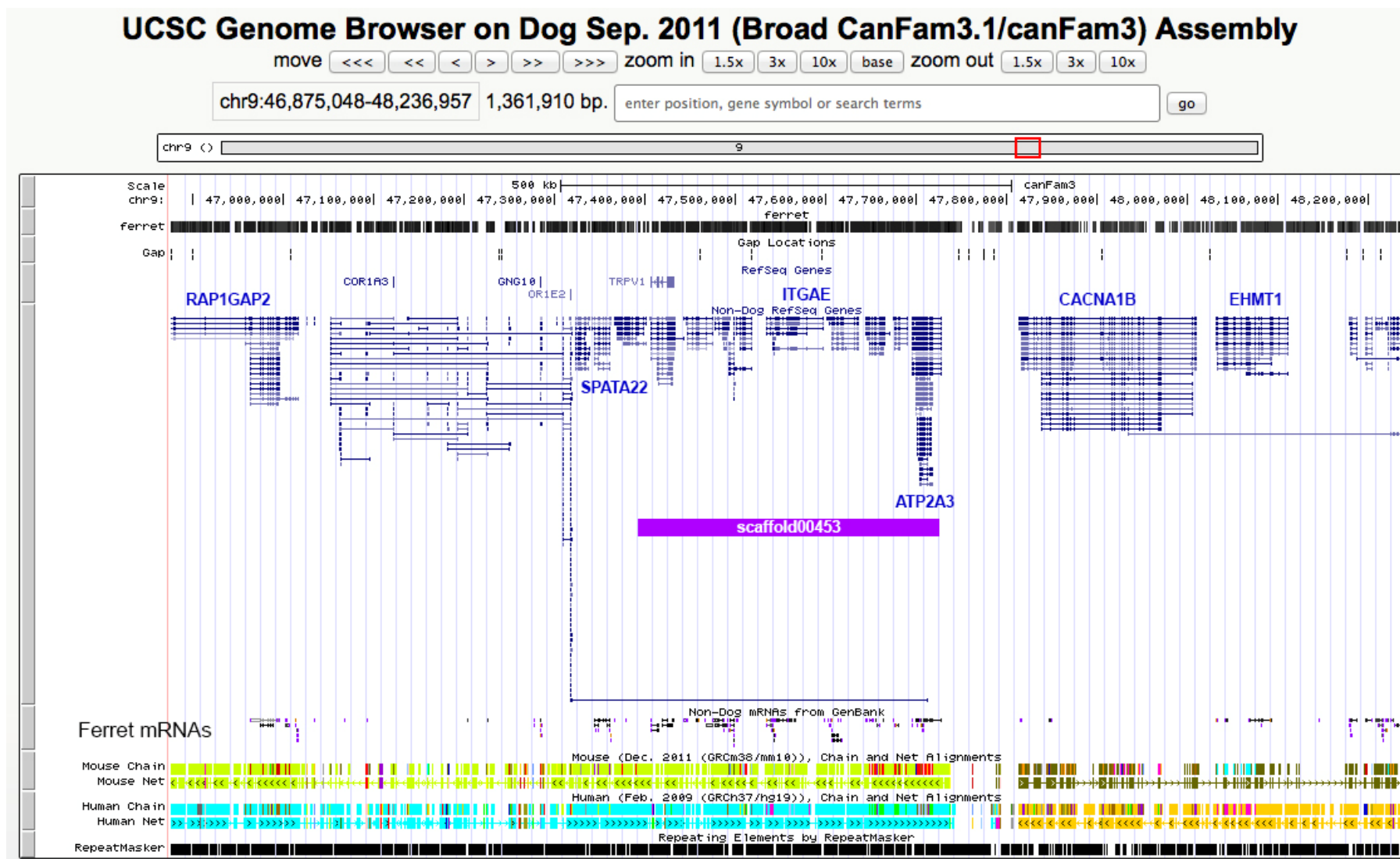
Supplementary Figure 17. Trachea exhibited much larger transcriptional changes than lung in influenza infected ferrets

Supplementary Figure 18. Workflow for directly comparing host transcriptional changes in trachea vs. lung of influenza infected ferrets

Supplementary Figure 19. The comparison of CF/non-CF gene expression changes in day 15 ferret vs. human bronchial brushing

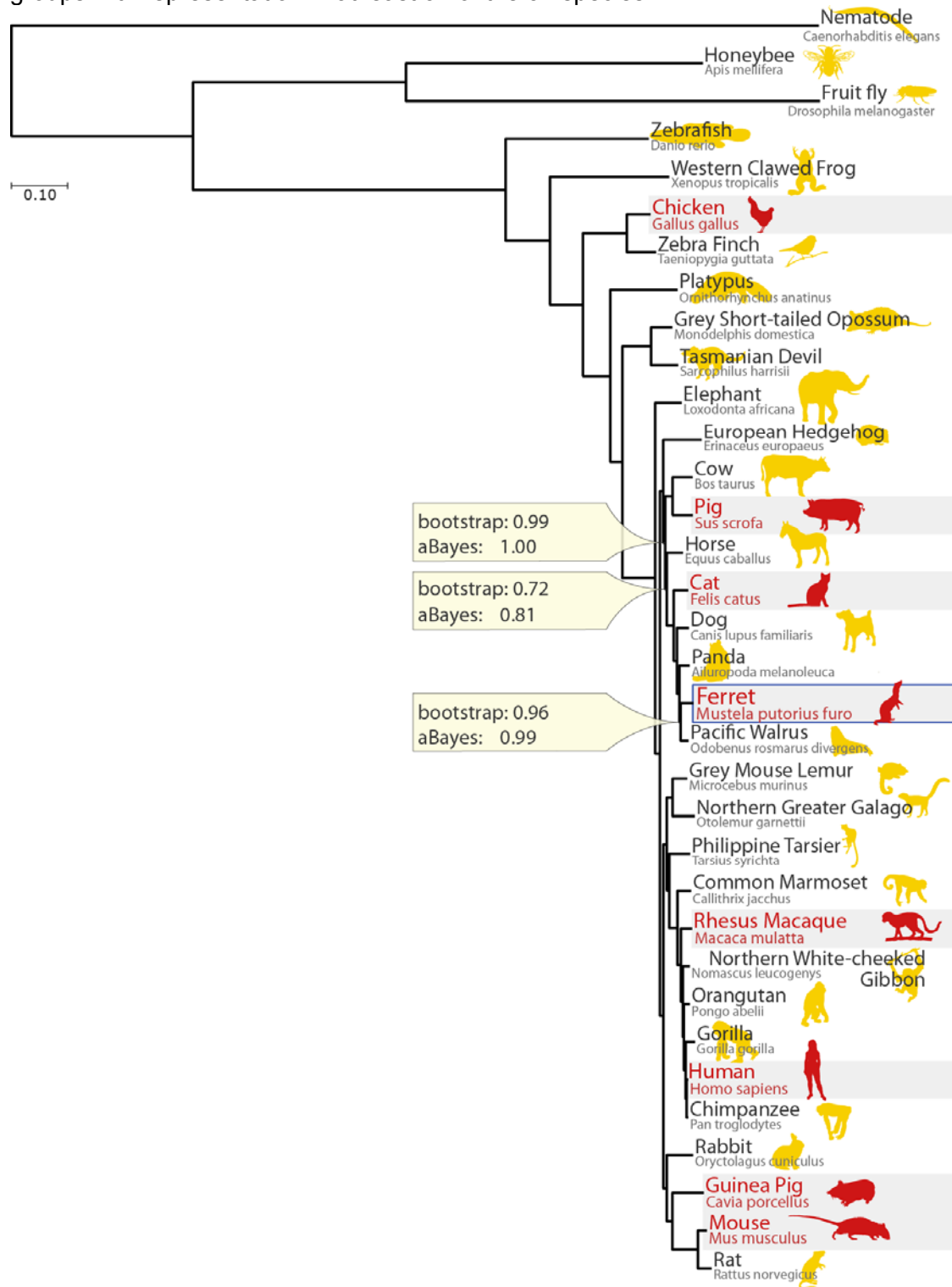
Supplementary Figure 20. Overview of genes differentially expressed in both day 15 CF ferret and CF human bronchial brushing

Supplementary Figure 1. View within the UCSC genome browser, showing a segment of dog chromosome 9 with the LiftOver of the ferret genome sequence indicated by the track labeled “ferret”.



The dog species was chosen as a close taxonomic neighbor with a chromosomally anchored genome assembly. Shown within the image is an alignment of ferret genome scaffold00453, one of the 7782 unplaced scaffolds of the ferret genome. Added labels have been placed on the condensed track of “Non-Dog RefSeq Genes” to illustrate the information content in this gene-rich area; this track provides information as to the probable locations of genes that are orthologous among many species. While this information would be present in browser view of just the ferret genome scaffold00453, the context of the LiftOver shows other genes that have a high probability of being on the same ferret chromosome as scaffold00453. Similarly, the genomic browser view of the LiftOver shows the syntenic context of ferret genome sequence vs. human and mouse chromosomes. Under the track of “Non-Dog mRNAs from GeneBank”, the browser window has been configured to show only mRNAs for *Mustela putorius furo*.

Supplementary Figure 2. Phylogenetic tree based on protein sequences of 789 orthologous groups with representation in at least 31 of the 34 species.



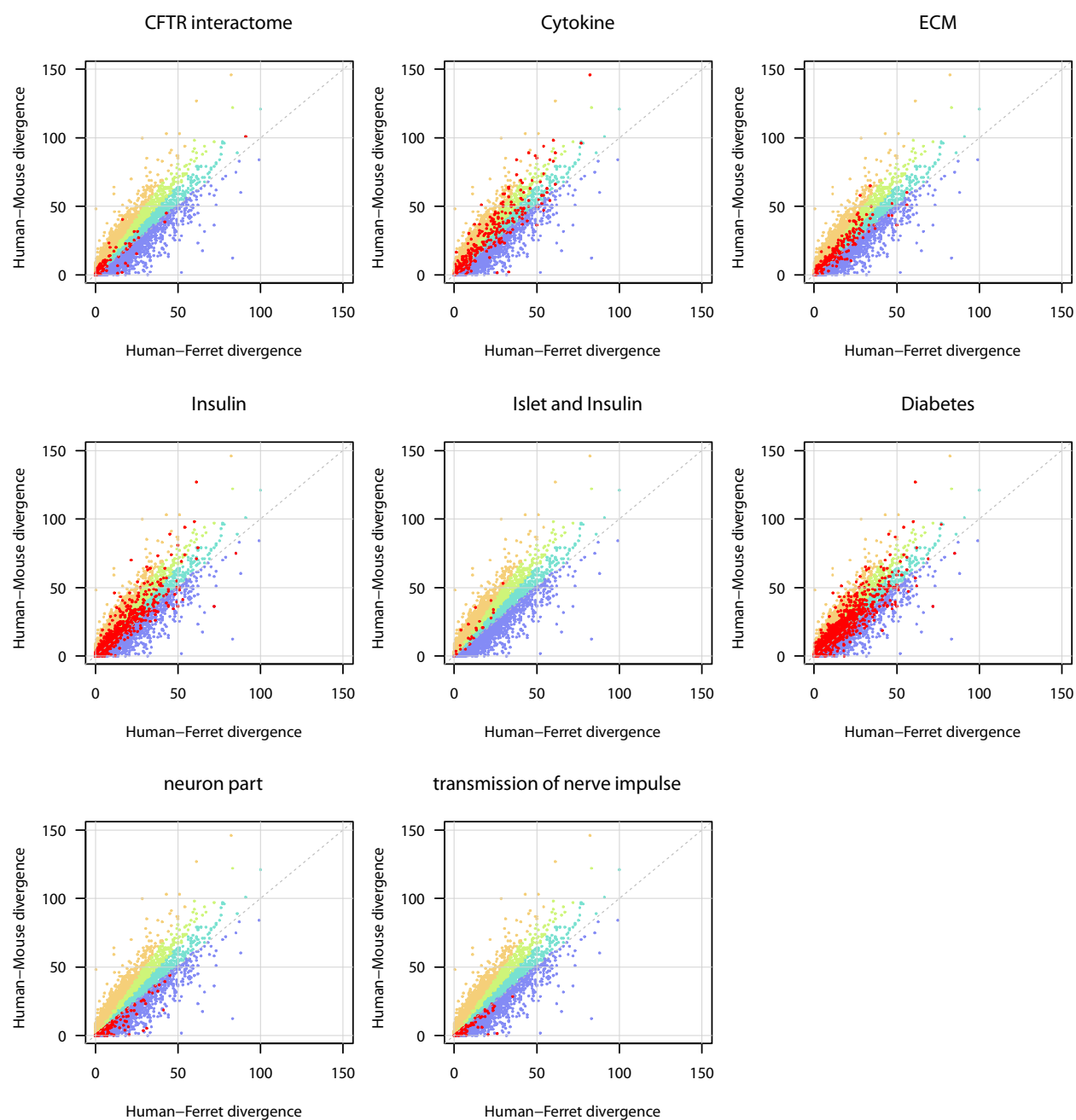
Species commonly used for respiratory models of human health and disease are highlighted in red. Bootstrap and approximate Bayesian support values are 1.0 except in the three branches indicated. Images are from www.phylopic.org.

Supplementary Figure 3. Gene Ontology enrichment for angular sector representing the top 25% of genes where the ferret sequence is closer to human than the mouse orthologue (see **Figure 1a** in main text).



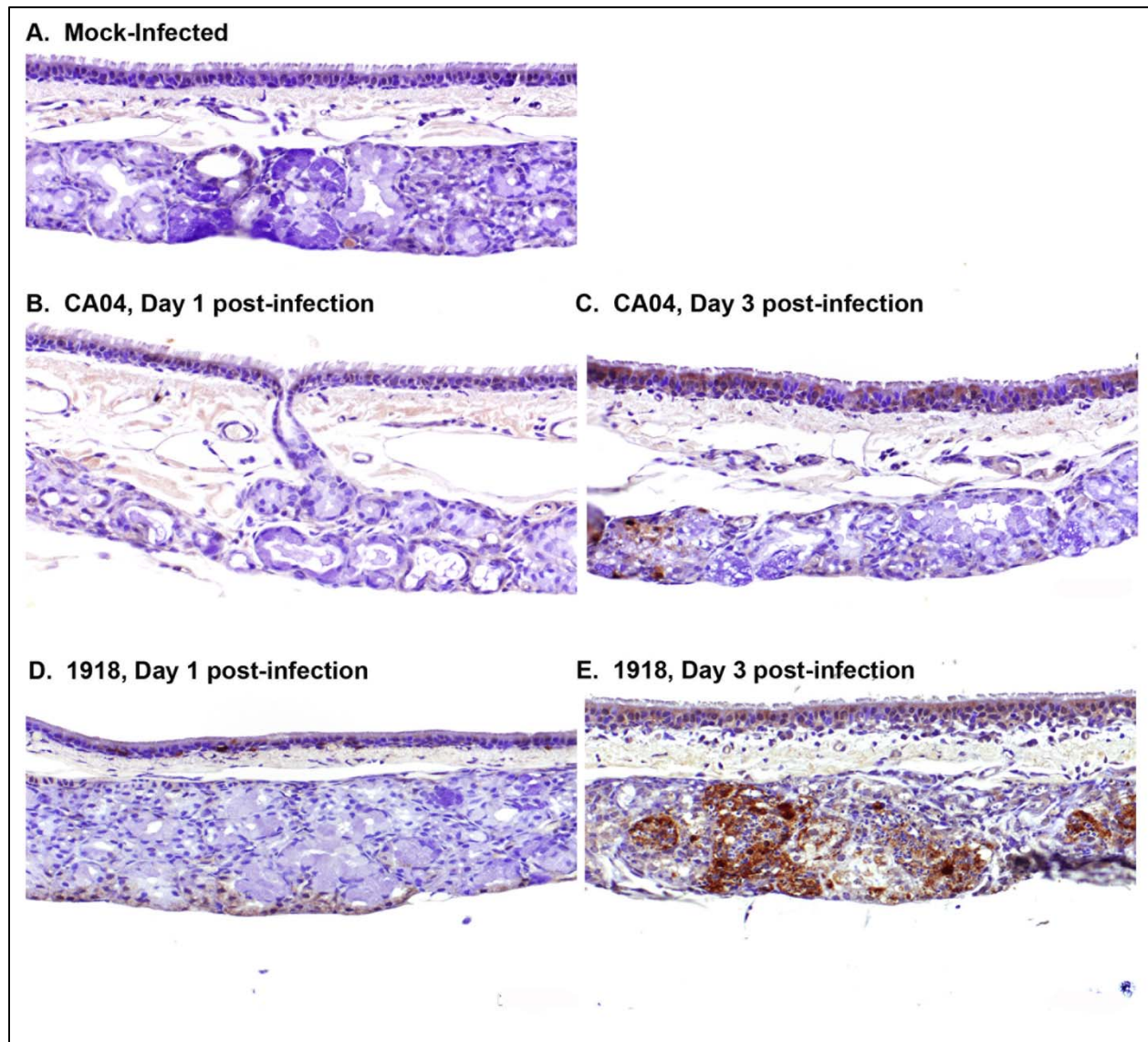
See Methods section for the description of how GO terms were clustered into the six groups shown in the figure. Headers to the boxed elements are summarizations of the clustered GO terms. Child terms are shown indented to the right below their parent. Using both Fisher's exact tests and Gene Set Enrichment Analysis (GSEA), we tested the enrichment of Gene Ontology (GO) terms for genes within the sectors of this plot, at a threshold false discovery rate q-value <0.05 .

Supplementary Figure 4. Protein sequence comparisons of orthologous trios (human, ferret, mouse); y axis: human vs. mouse distance; x axis: human vs. ferret distance.



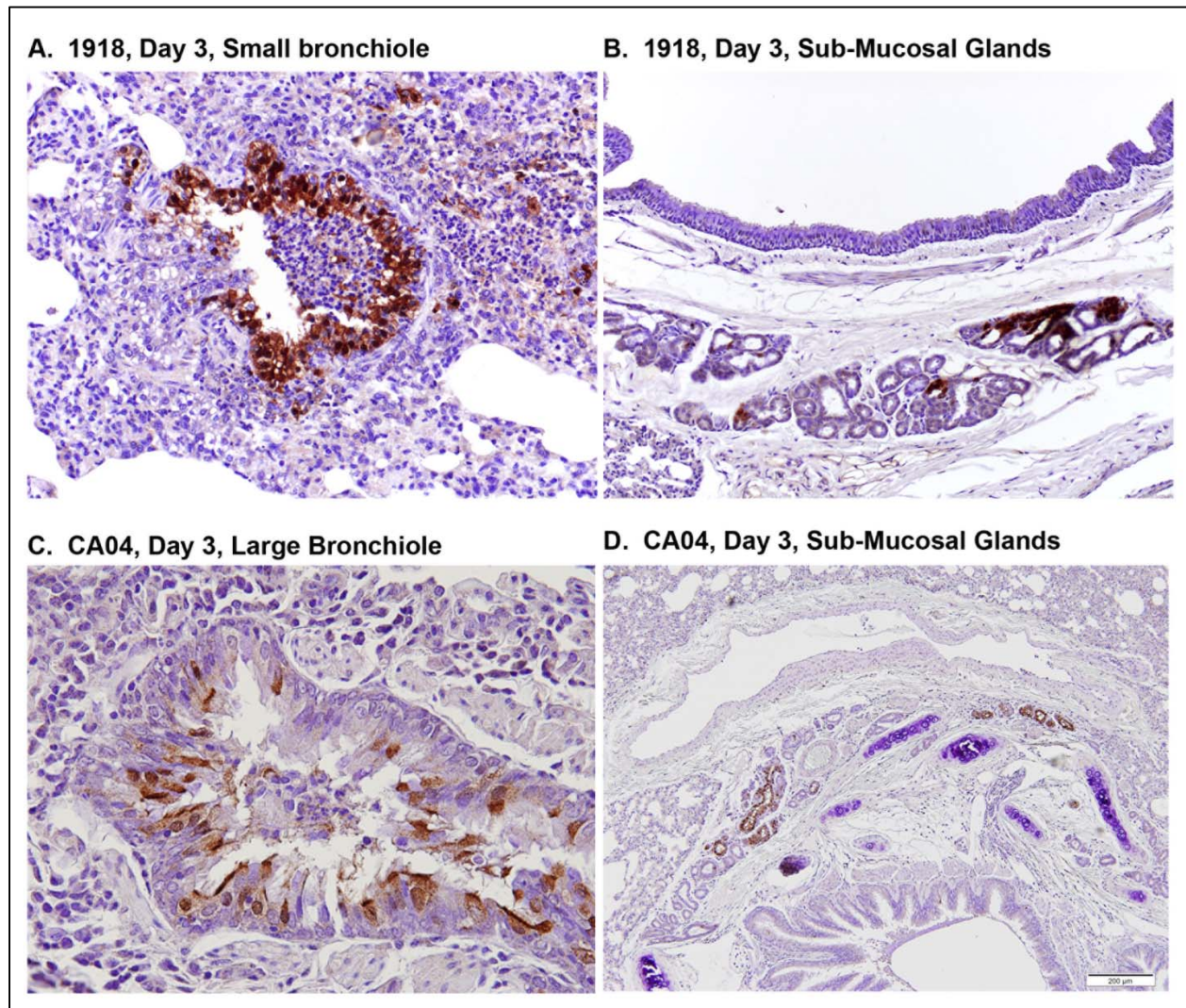
Red points correspond to the gene subset associated with the category shown in the panel title.

Supplementary Figure 5. Ferret tracheal immunohistochemistry.



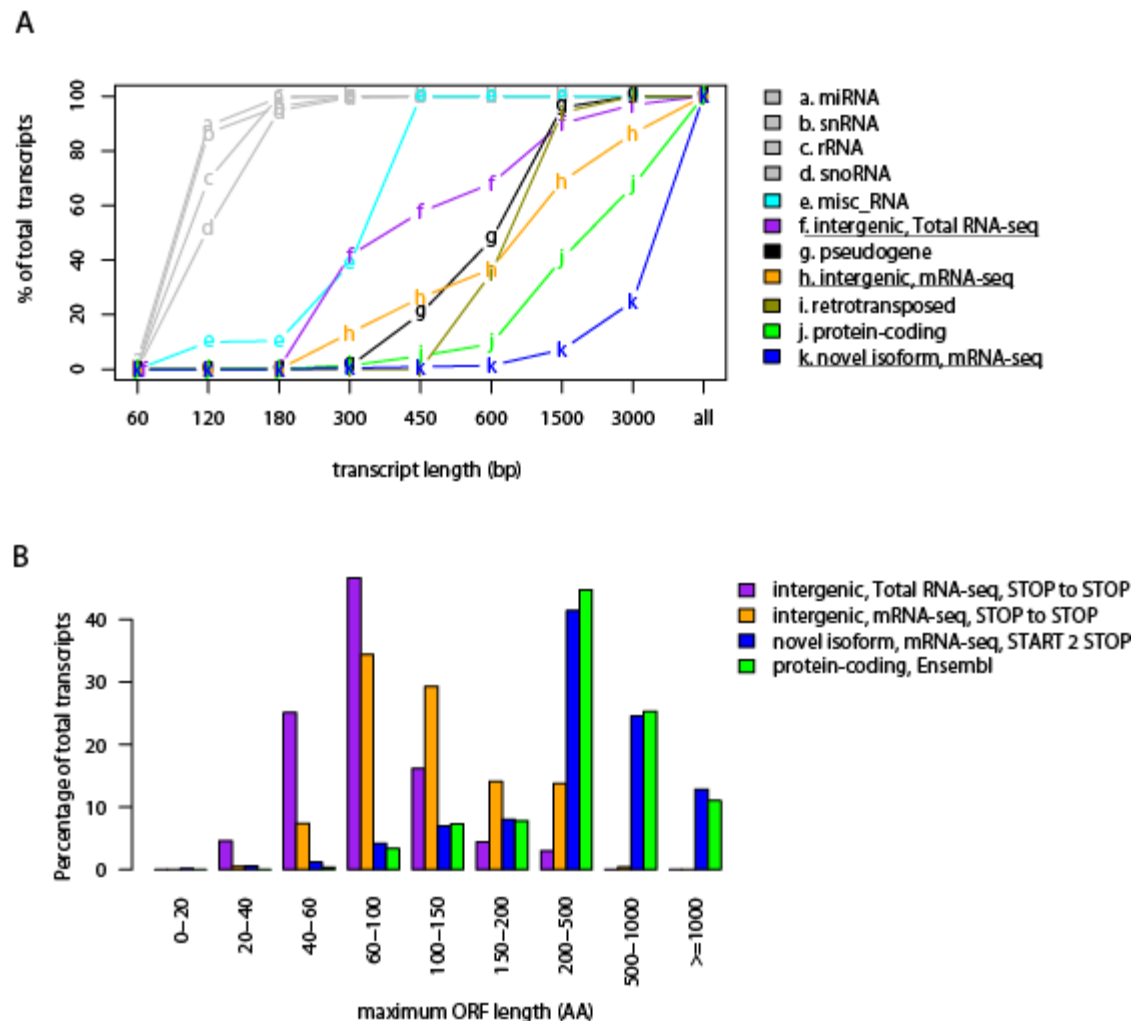
Sections of ferret trachea were stained with hematoxylin, and counterstained with a polyclonal anti-influenza antibody, as described in the Methods section. Representative images for mock-infected ferrets (A), CA04-infected ferrets at days 1 and 3 post-infection (B and C, respectively), and 1918-infected ferrets at days 1 and 3 post-infection (D and E, respectively) are shown. Brown staining indicates tissues that are positive for influenza virus antigens.

Supplementary Figure 6. Ferret lung immunohistochemistry.



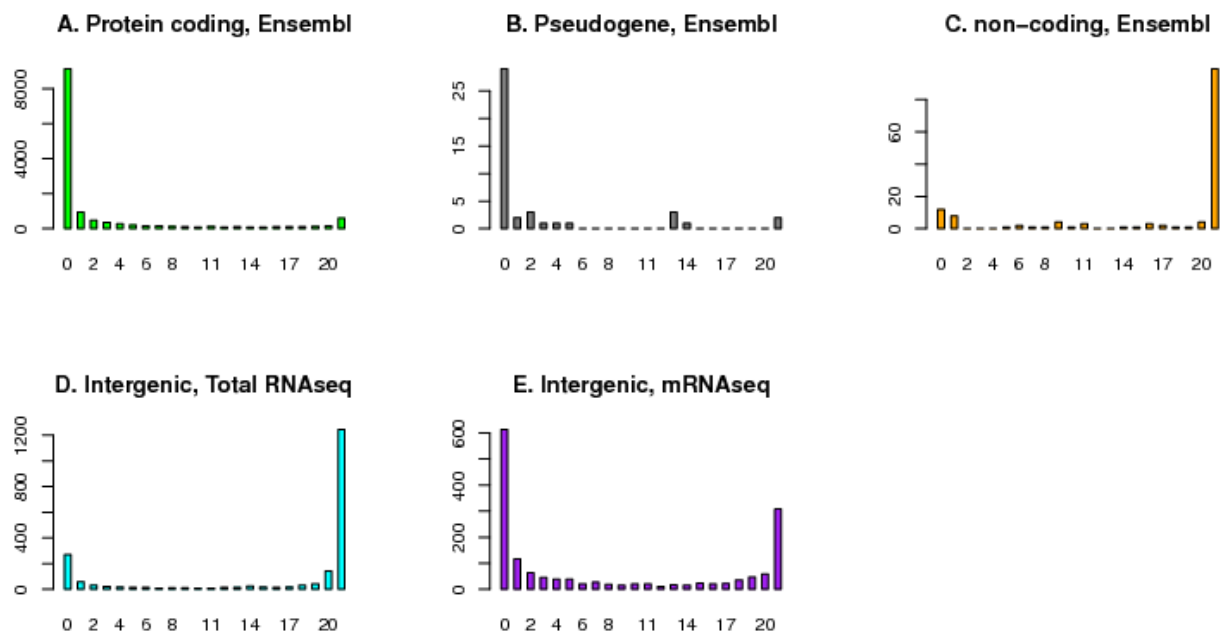
Sections of ferret lung were stained with hematoxylin, and counterstained with a polyclonal anti-influenza antibody, as described in the Methods section. Representative images of ferret bronchioles and sub-mucosal glands from 1918–infected ferrets (A and B) and CA04–infected ferrets (C and D) at day 3 post-infection are shown. Brown staining indicates tissues that are positive for influenza virus antigens.

Supplementary Figure 7. Characterization of intergenic transcripts derived from ferret RNA-seq data.



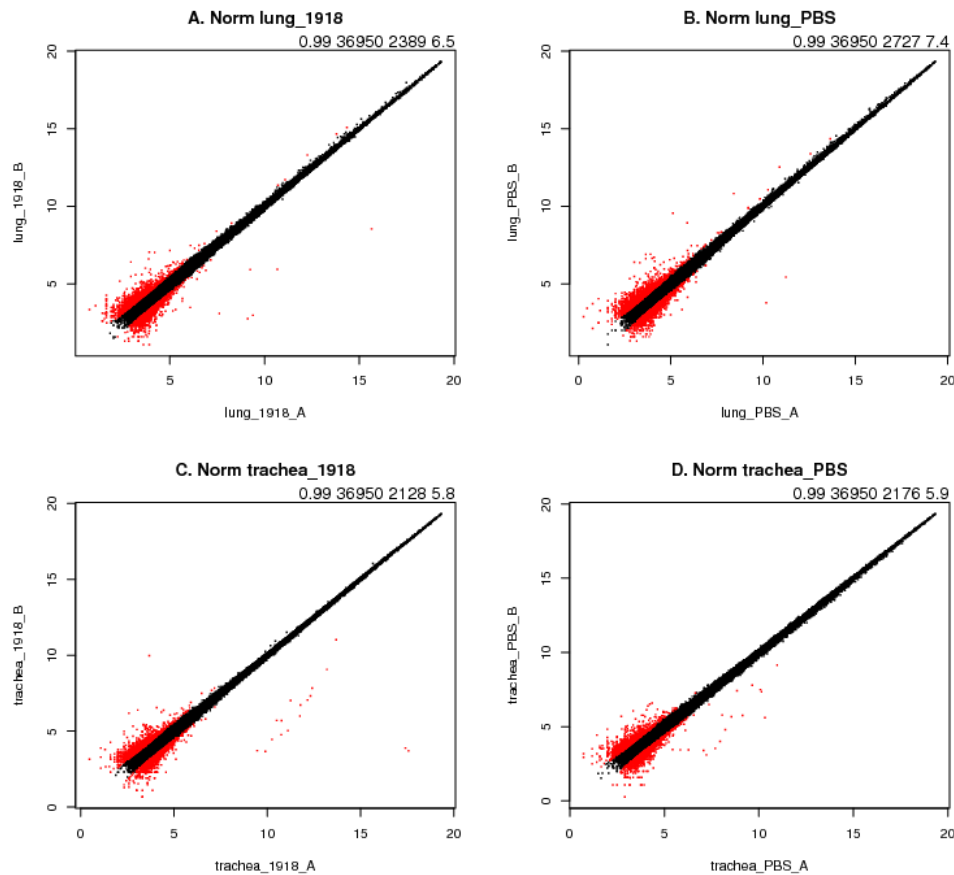
A) Length distribution of un-annotated transcripts and annotated transcripts, showing that the un-annotated intergenic transcripts (underlined; f. and h.) are longer than Ensembl annotated small RNAs (gray; a., b., and d.). **B)** Length distribution of longest open reading frames (ORFs) of un-annotated transcripts. For intergenic transcript, the maximum ORF length of six-frame stop to stop translation was plotted to assess coding potential. For predicted novel isoforms of Ensembl annotated protein-coding genes, the maximum ORF length of sense three-frame start to stop translation was plotted to assess if they cover full-length proteins.

Supplementary Figure 8. Comparison of transcript abundances measured by Total RNA-seq vs. mRNA-seq analysis of the same set of ferret lung samples.



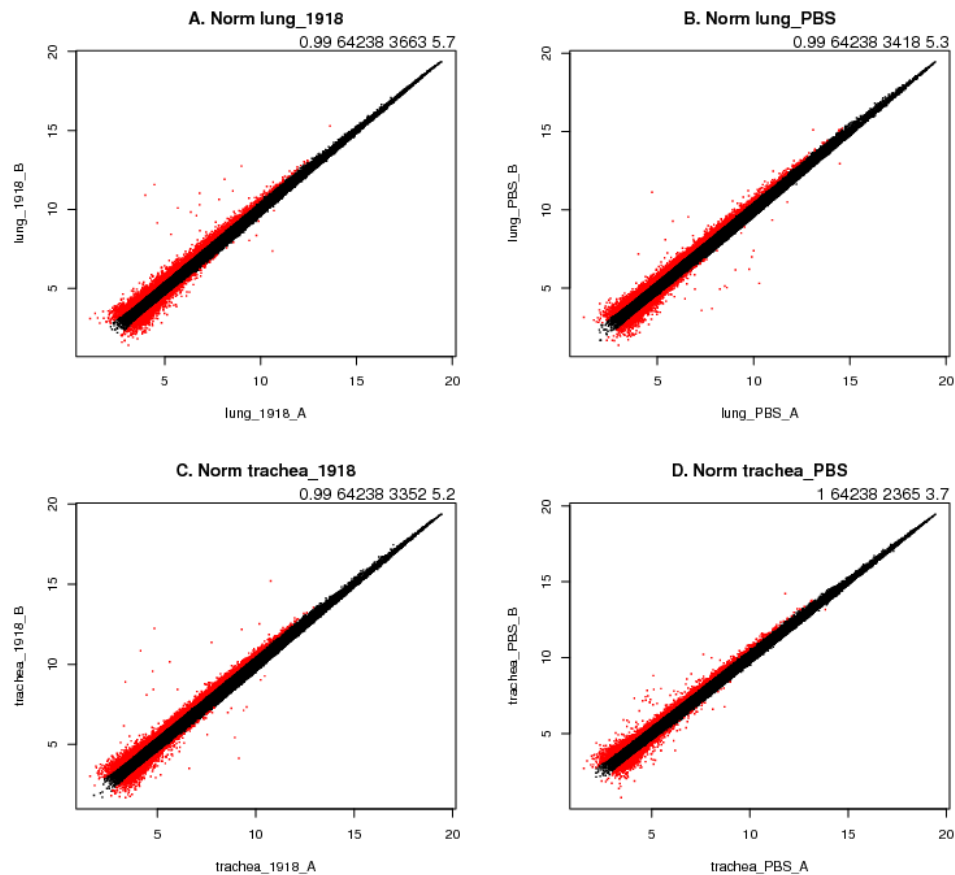
A) Ensembl annotated protein-coding genes. Each bar represent the number of ferret lung samples (21 in total) as labeled below in which the number of genes (the height of each) had more reads in Total RNA-seq analysis than mRNA-seq analysis of the same samples. **B)** Ensembl annotated pseudogenes. **C)** Ensembl annotated non-coding RNAs. **D)** Intergenic regions from Total RNA-seq analysis. **E)** Intergenic regions from mRNA-seq analysis.

Supplementary Figure 9. Comparison of microarray technical replicates (ferret microarray version 1).



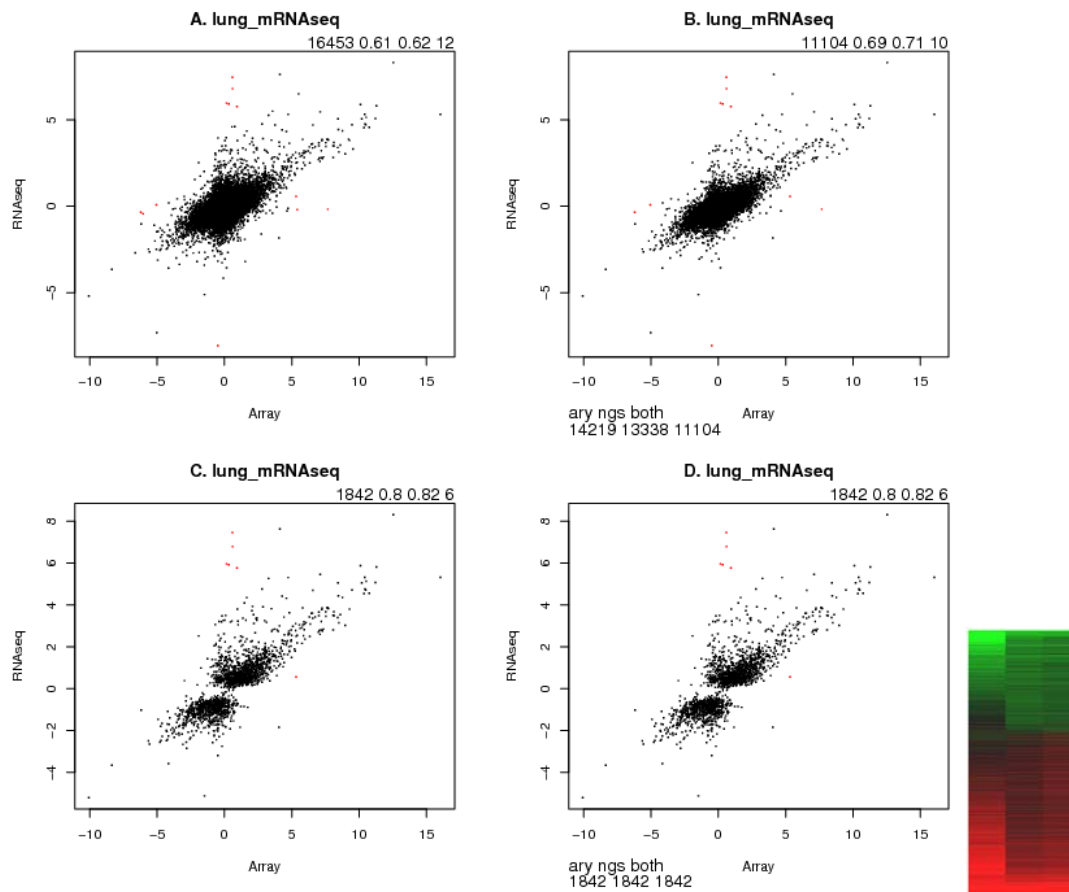
Each scatterplot represents two technical replicates of one biological sample. X-axis is gene probe intensities (replicated probes were averaged, log2 scale) from one technical replicate and y-axis is the intensities of the corresponding genes from another technical replicate. Colored in red is the subset of genes with intensity difference between two technical replicates greater than 1.5 fold. The Pearson correlation between technical replicates, the total number of genes compared, the number of genes in red, and the percentage of genes in red are shown in the top right corner of each plot.

Supplementary Figure 10. Comparison of microarray technical replicates (ferret microarray version 2).



Similar as in Supplementary Figure 9: Each scatterplot represents two technical replicates of one biological sample. X-axis is gene probe intensities (replicated probes were averaged, log2 scale) from one technical replicate and y-axis is the intensities of the corresponding genes from another technical replicate. Colored in red is the subset of genes with intensity difference between two technical replicates greater than 1.5 fold. The Pearson correlation between technical replicates, the total number of genes compared, the number of genes in red, and the percentage of genes in red are shown in the top right corner of each plot.

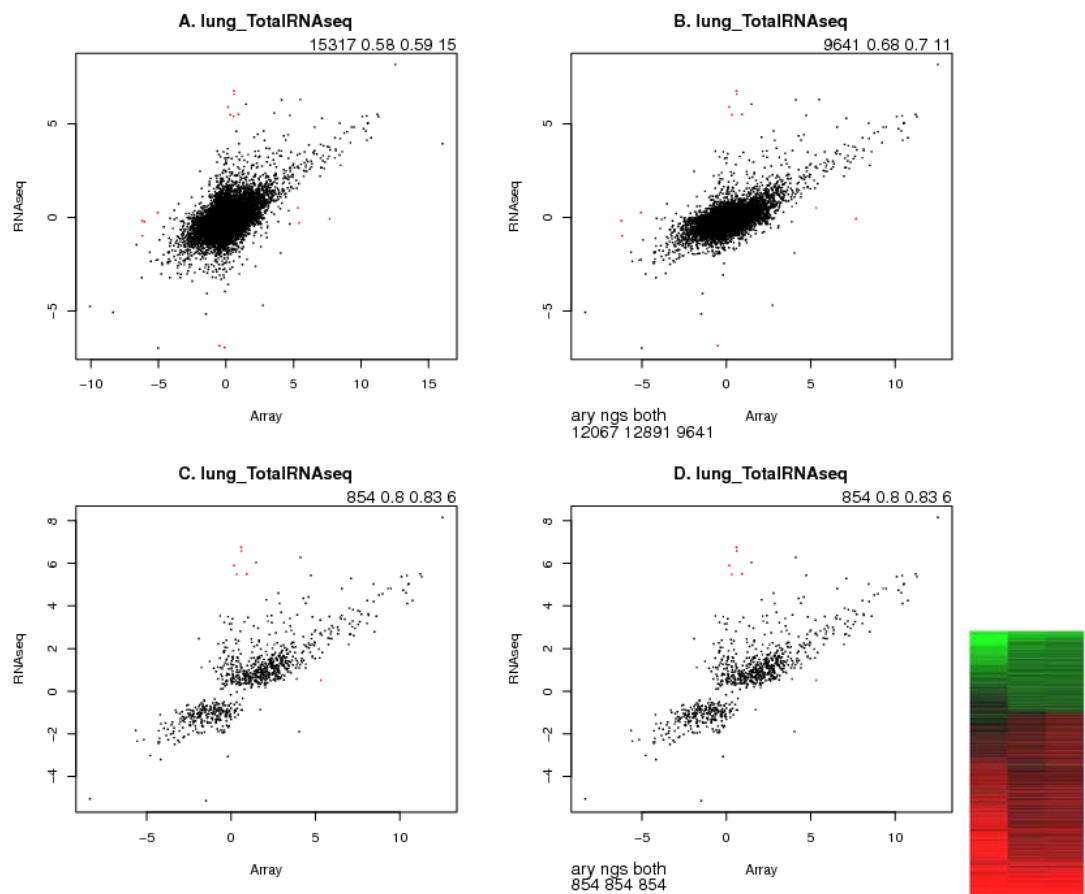
Supplementary Figure 11. Ferret microarray (version 1) vs. RNA-seq comparison (lung mRNA-seq).



A) The scatterplot of infection/mock log₂ fold changes by microarray (X-axis) vs. those of corresponding genes measured by RNA-seq (Y-axis), on the same pair of samples. For microarray replicated probes were averaged first. Then two technical replicates were averaged before the calculation of infection/mock fold changes. Only genes used for original RNA-seq differential analysis, i.e., those genes with a raw read count of 10 or more in at least 3 out of 21 samples, are included here. The numbers in the top right corner of the plot are: the total number of genes compared, the overall Pearson correlation, the Pearson correlation after the removal of genes showing obviously large discrepancies (colored in red), i.e. those with absolute log₂ fold change less than 1 by one measurement but greater than 5 by another measurement, and the number of genes in red. **B)** Same as in **A**, but with the removal of a subset of genes likely of low

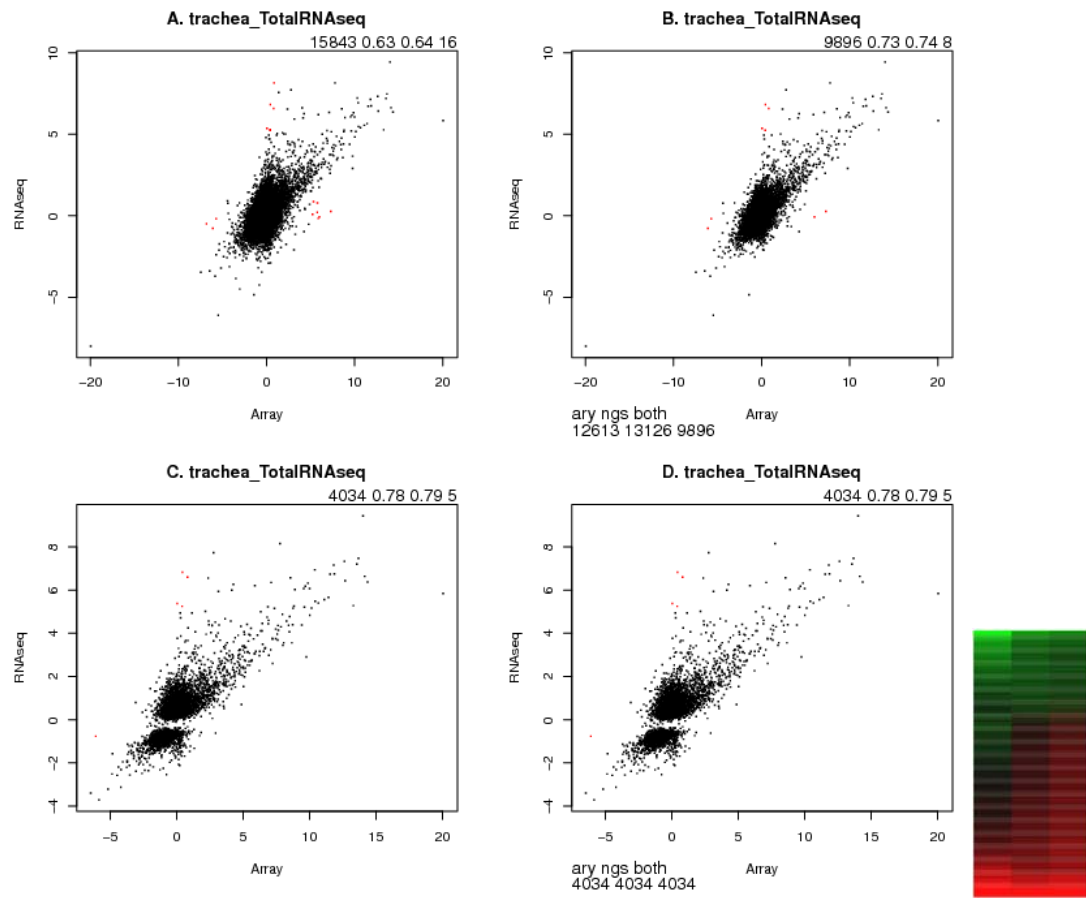
quality measurements by either method. Specifically, for microarray the genes with 1.5 fold or greater difference between technical replicates or the log2 intensities in both biological samples smaller than 5 were filtered out. For RNA-seq, genes with raw read count less than 50 in both biological samples were filtered out. The number of genes remained by either filter or both filters are shown at the bottom left corner. **C)** similar as in **A**, only genes which were identified as differentially expressed (adjusted p-value < 0.01) between 1918 virus infections (day 1 after infection) and mock infections by the original mRNA-seq analysis. **D)** similar in **C**, with same filtering as applied in **B**. (In this case there were no genes filtered out.) The heatmap gives another view of log 2 infection/mock fold changes for genes shown in **D**, columns are in the order as: microarray, RNA-seq, the average ratios of three replicates by RNA-seq.

Supplementary Figure 12. Ferret microarray (version 1) vs. RNA-seq comparison (lung Total RNA-seq).



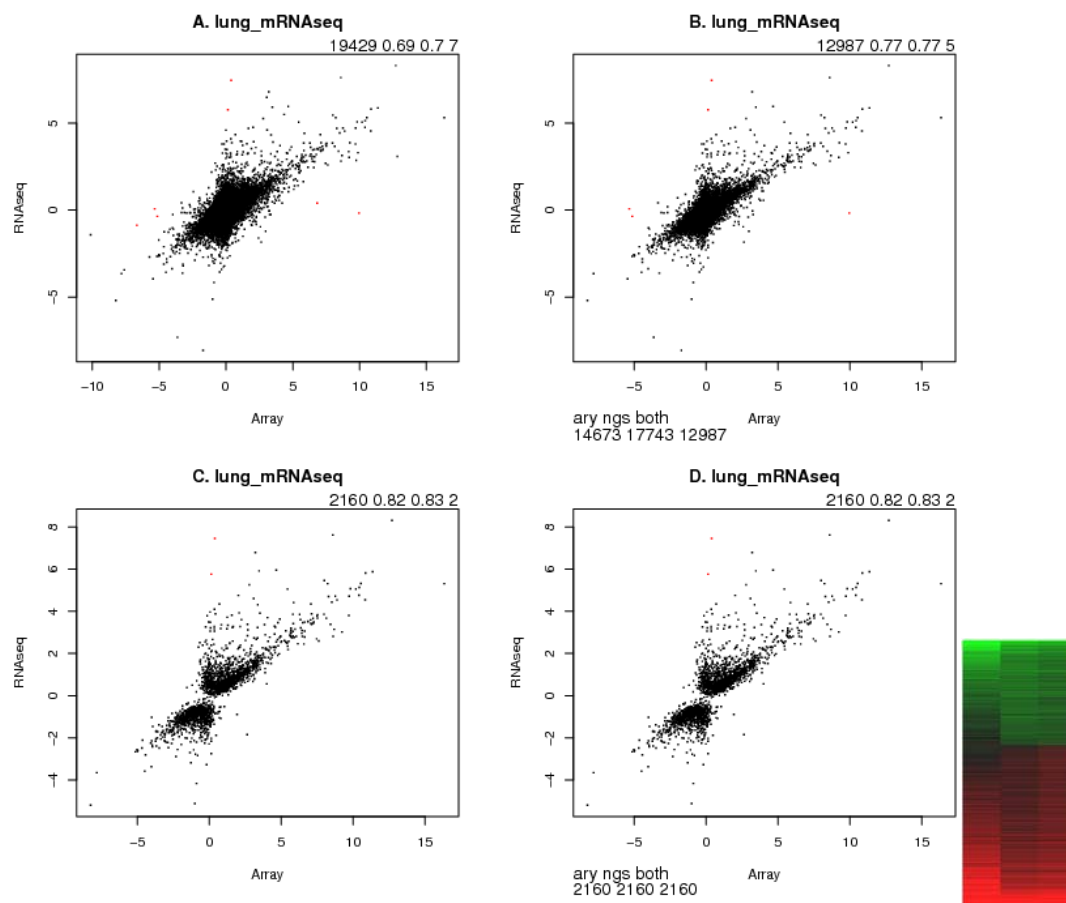
Similar as **Supplementary Figure 11.**

Supplementary Figure 13. Ferret microarray (version 1) vs. RNA-seq comparison (trachea Total RNA-seq).



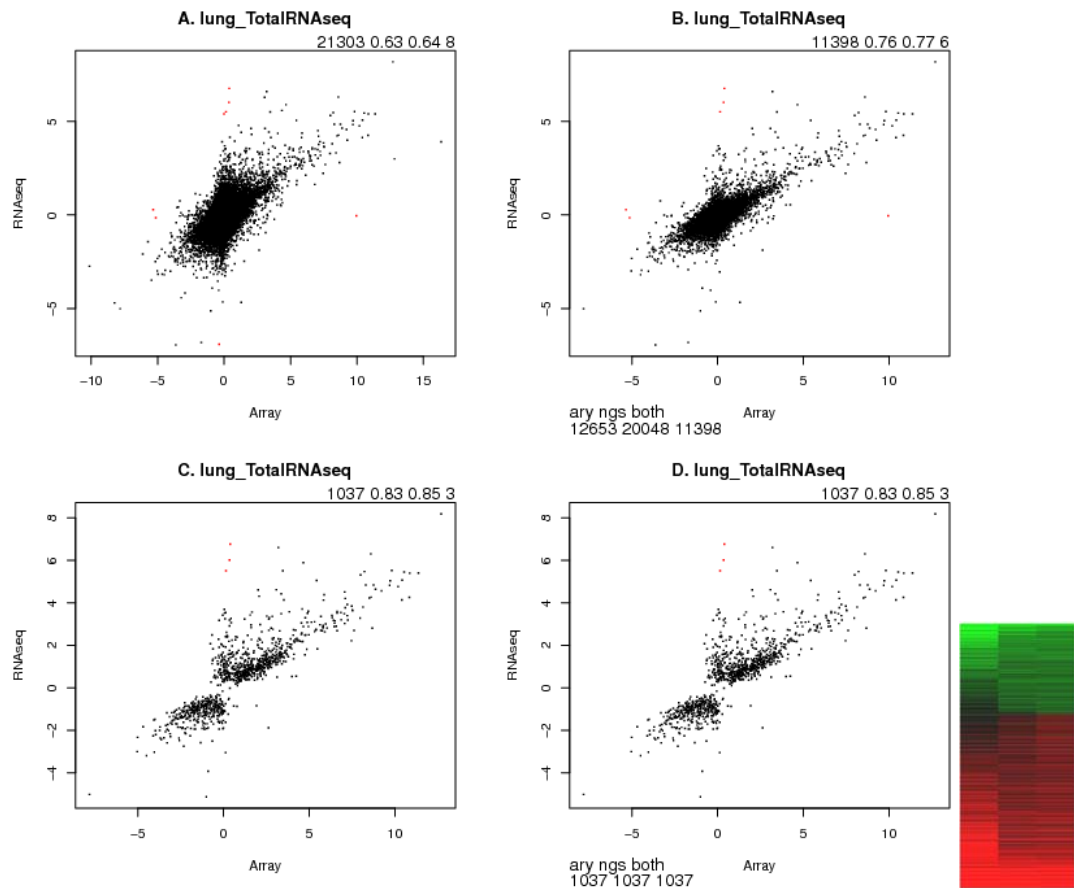
Similar as **Supplementary Figure 11.**

Supplementary Figure 14. Ferret microarray (version 2) vs. RNA-seq comparison (lung mRNA-seq).



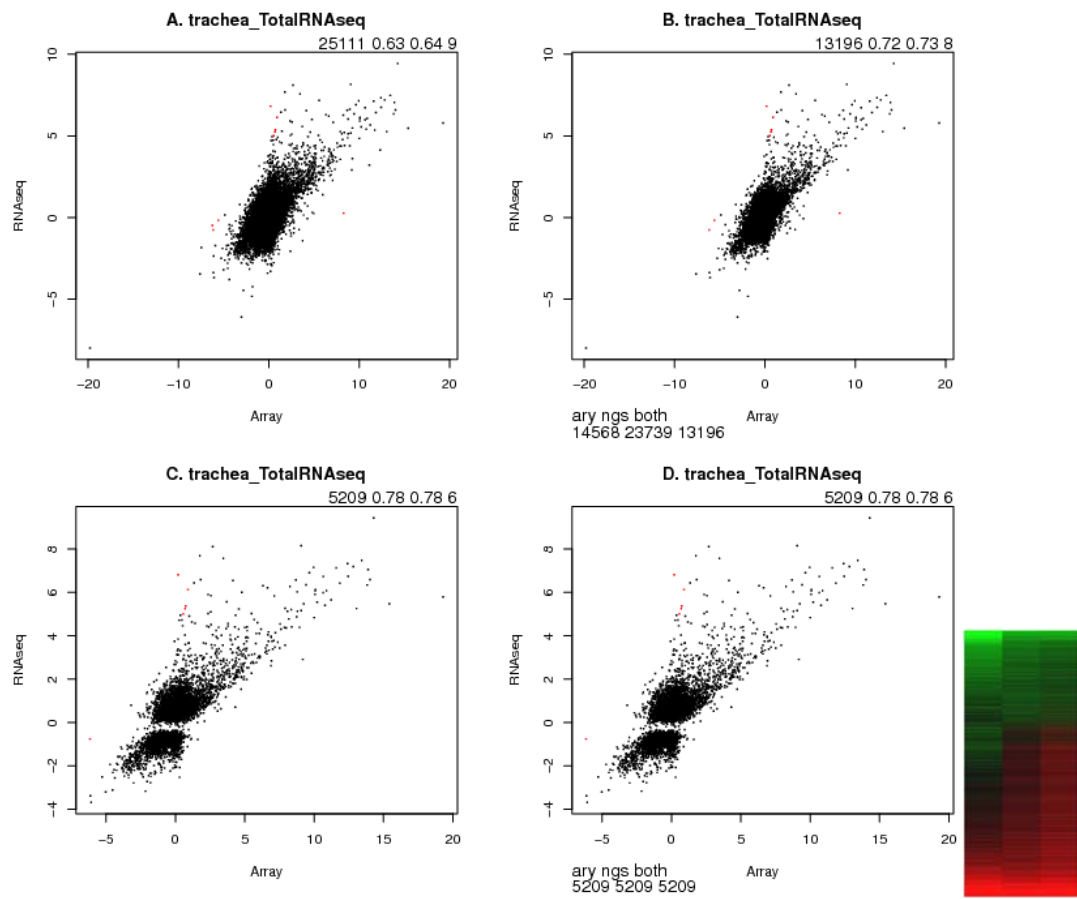
Similar as **Supplementary Figure 11**.

Supplementary Figure 15. Ferret microarray (version 2) vs. RNA-seq comparison (lung Total RNA-seq).



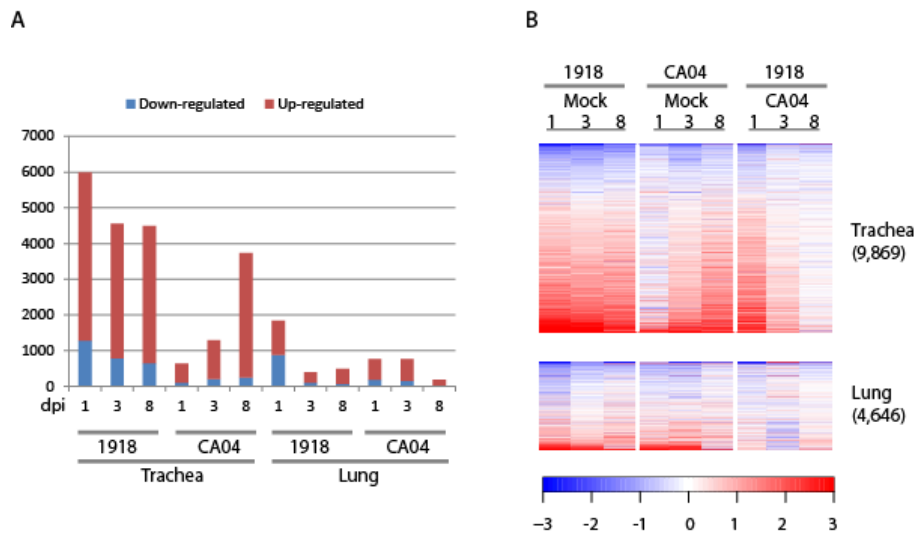
Similar as **Supplementary Figure 11.**

Supplementary Figure 16. Ferret microarray (version 2) vs. RNA-seq comparison (trachea Total RNA-seq).



Similar as **Supplementary Figure 11**.

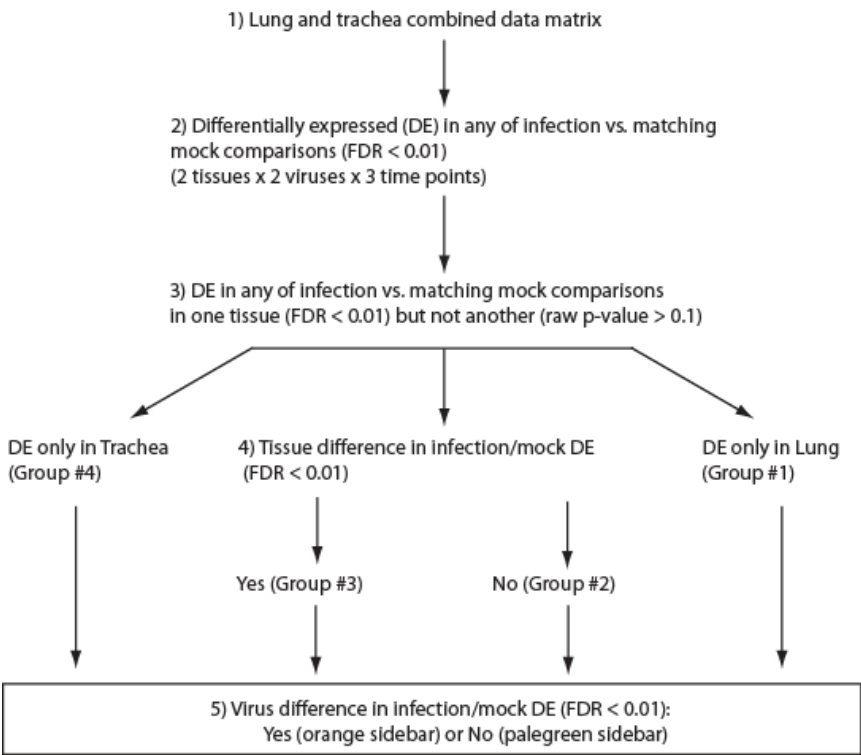
Supplementary Figure 17. Trachea exhibited much larger transcriptional changes than lung in influenza infected ferrets.



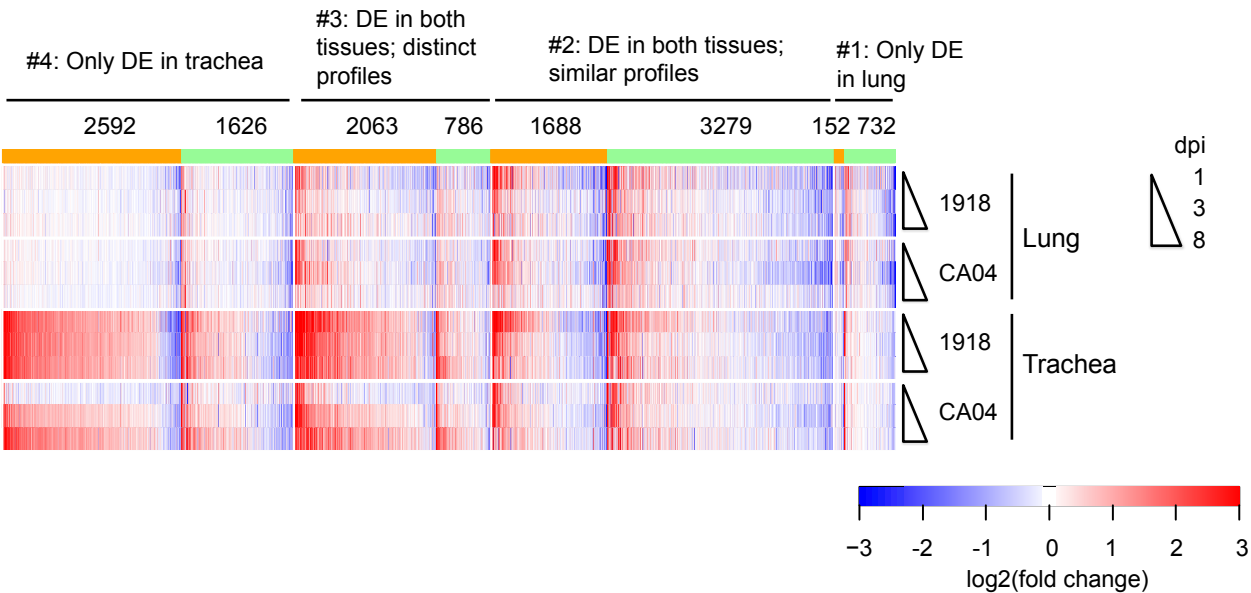
A) The number of differentially expressed (adjusted p-value < 0.01) ferret genes in each infection vs. mock comparisons within each tissue. Lung and trachea Total RNA-seq data was analyzed together as shown in Supplementary Figure 18. **B)** Overview of expression changes of ferret genes differentially expressed in any of the infection vs. mock comparisons (adjusted p-value < 0.01) within each tissue. The heatmap shows log2 infection vs. mock fold changes, rows are genes and columns are infection conditions.

Supplementary Figure 18. Workflow for directly comparing host transcriptional changes in trachea vs. lung of influenza infected ferrets.

A



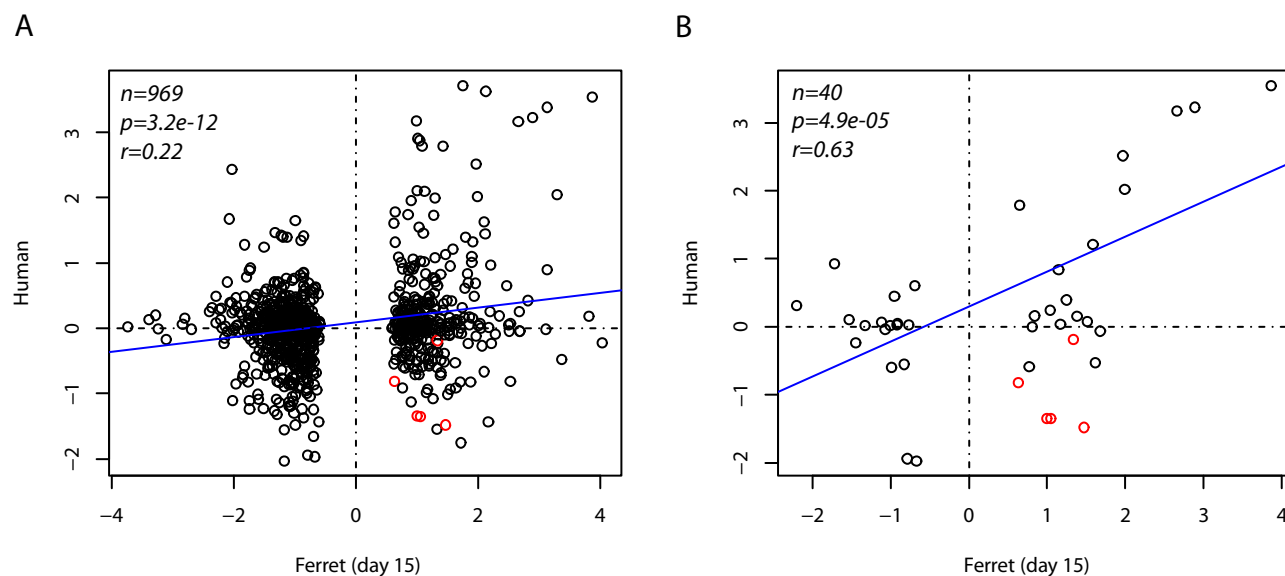
B



A) The diagram of the workflow. The expression data from both tissues were combined and processed together, using the generalized linear model approach provided by edgeR (Methods).

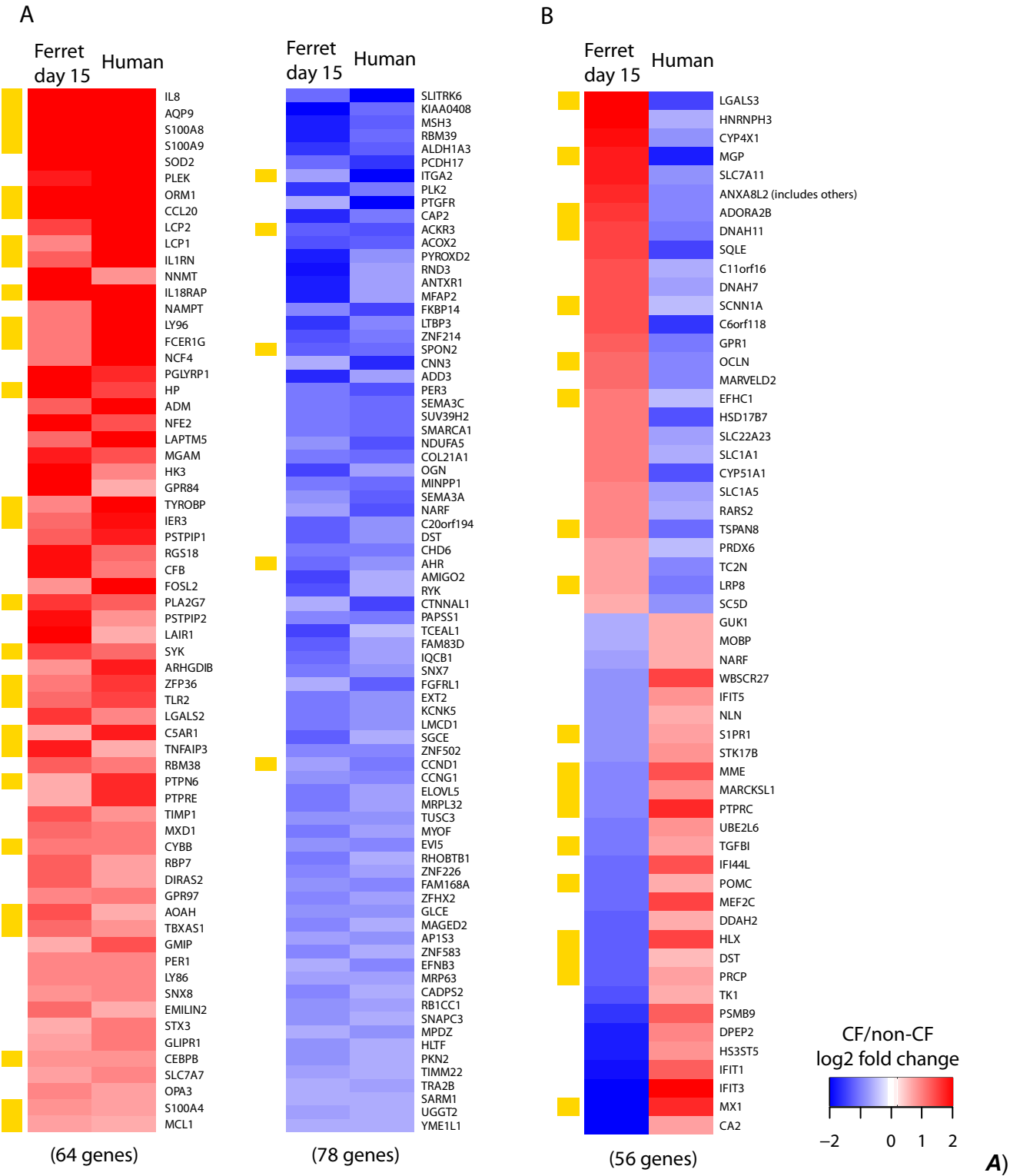
Step 2 defined the set of infection related ferret genes, which were differentially expressed in any of the infection conditions compared to matched mocks. Steps 3-5 partitioned this set of infection related genes into disjointed clusters in a hierarchical order. Step 3 singled out those genes which were significantly differently expressed in one tissue but not another. The remaining genes were further partitioned into two groups by statistically evaluating if their transcriptional changes were different between two tissues (Step 4). Lastly within each group, genes showing significant differences between two virus infections or not were identified (Step 5). To illustrate the overall tissue differences in transcriptional changes, the differentially expressed genes identified for each tissue as described above were extracted, and visualized in Supplementary Figure 17. **B)** A more detailed representation of gene clustering. Orange subsections indicate genes that differ between the virus strains, green subsections indicate genes that do not (Step 5 from part **a**).

Supplementary Figure 19. The comparison of CF/non-CF gene expression changes in day 15 ferret vs. human bronchial brushing.



A) Scatterplot of log2 CF/non-CF fold changes in day 15 ferrets (x-axis) vs. that in human (y-axis) for those genes differentially expressed in day 15 CF ferrets. Only genes with homologs mapped between two species and profiled on both microarray platforms were included. The numbers in the upper left corner are: the number of gene pairs plotted (n), the ANOVA p -value (p) of the fitted linear regression model (blue line in the plot), and Pearson correlation coefficient (r). Colored in red are five genes from the Cholesterol Biosynthesis pathway by IPA annotation. **B)** Similarly as in A, but only for the union of ferret genes in four canonical pathways shown in Supplementary Table 17.

Supplementary Figure 20. Overview of genes differentially expressed in both day 15 CF ferret and CF human bronchial brushing.



Heatmap of the subset of genes with the expressions changed in the same direction between ferret and

human datasets, up-regulated in the CF vs. non-CF comparison on the left and down-regulated on the right. On the left side of the heatmaps, genes of the IPA function 'inflammatory response' are marked in gold. The 'inflammatory response' was returned with an enrichment p-value of $2.8\text{e-}14$ when the IPA functional analysis was performed on these 142 genes. **B)** Similarly as in **A**, but for genes with the CF vs. non-CF expression changes in the opposite direction between ferret and human datasets. The marked genes were from the IPA function 'cell movement', which was returned with an enrichment p-value of $2.11\text{e-}05$ when the IPA functional analysis was performed on these 56 genes.

II. Supplementary Tables

Supplementary Table 1: Tissue RNAseq data for annotation **(Excel file upload separately)**

Supplementary Table 2: Summary of species used in the construction of phylogenetic tree **(Excel file upload separately)**

Supplementary Table 3: Protein alignments used for the generation of phylogenetic tree **(Text file upload separately)**

Supplementary Table 4: Human-mouse and human-ferret protein divergences in PAM metric **(Excel file upload separately)**

Supplementary Table 5: Gene Ontology enrichment analysis by Gene Set Enrichment Analysis (GSEA) for the angular sector representing the top 25% of genes where the ferret sequence is closer to human than the mouse orthologue **(Excel file upload separately)**

Supplementary Table 6: Gene Ontology enrichment analysis by Fisher's exact tests for the angular sector representing the top 25% of genes where the ferret sequence is closer to human than the mouse orthologue **(Excel file upload separately)**

Supplementary Table 7: Concordance of tissue abundant genes between ferret and human

Supplementary Table 8: Assignment of tissue specific ferret-human gene pairs by clustering **(Excel file upload separately)**

Supplementary Table 9: Comparison of the overall abundance of gene clusters across tissues of the same species

Supplementary Table 10: Ferret-human sequence divergence analysis of transcription factors in tissue abundant clusters

Supplementary Table 11: PAM metrics (ferret to human) evolutionary distance, tissue-specific transcription factors **(Excel file upload separately)**

Supplementary Table 12: Virus titer and viral qPCR measurements of influenza infected ferret lung and trachea samples **(Excel file upload separately)**

Supplementary Table 13: Summary of the expanded ferret annotations derived from RNA-seq data

Supplementary Table 14: Summary of ferret microarray design

Supplementary Table 15: Genes differentially expressed in trachea and lung samples from influenza infected ferrets **(Excel file upload separately)**

Supplementary Table 16: Functional enrichment analysis of lung-specific and tracheal-specific DE genes that show statistically significant differences between 1918 and CA04

Supplementary Table 17: Top four significant IPA canonical pathways between newborn ferrets and 15 day ferret vs adult human lung samples **(Excel file upload separately)**

Supplementary Table 18: Top three significant IPA disease functional pathways between newborn ferrets and 15 day ferret vs adult human lung samples **(Excel file upload separately)**

Supplementary Table 7. Concordance of tissue abundant genes between ferret and human.

Tissue	Number of genes with highest relative abundance			P-value	Expected
	Ferret	Human	Both		
Brain	2868	2176	1561	0	494
Testis	3360	3946	2253	0	1049
Skeletal muscle	813	1011	362	0	65
Heart	886	928	284	4.29E-187	65
Lung	2219	1728	928	0	303
Liver	763	841	367	0	51
Kidney	1727	2006	730	1.96E-228	274

Supplementary Table 9. Comparison of the overall abundance of gene clusters across tissues of the same species.

Cluster	Tissue	Ferret							Human						
		Brain	Testis	Muscle	Heart	Lung	Liver	Kidney	Brain	Testis	Muscle	Heart	Lung	Liver	Kidney
1	Brain		0	0	0	0	0	0		0	0	0	0	0	0
2	Testis	0		0	0	0	0	0	0		0	0	0	0	0
3	Skeletal muscle	2.54E-163	3.07E-179		3.13E-28	7.12E-174	9.96E-193	3.44E-177	1.20E-132	1.59E-115		7.01E-06	4.56E-141	6.05E-166	1.43E-129
3	Heart	2.71E-102	6.09E-131	1		2.62E-115	3.68E-150	2.49E-118	4.61E-105	2.19E-85	0.999993		6.52E-114	1.26E-143	1.29E-101
4	Lung	0	0	0	0		0	0	0	0	0	0		0	0
5	Liver	5.28E-129	6.87E-127	2.32E-130	3.90E-128	3.80E-124		1.15E-103	4.85E-129	1.95E-127	3.51E-132	3.83E-130	9.53E-129		3.62E-119
6	Kidney	6.53E-167	2.01E-163	5.95E-169	1.58E-169	8.53E-158	3.04E-161		2.98E-150	1.48E-130	3.43E-163	4.98E-160	1.87E-129	1.89E-116	

For each cluster of genes as shown in **Figure 1c**, the overall relative abundance in one tissue was compared to that of all other tissues of the same species using Mann-Whitney two-sample test. The table shows the p-values with alternative hypothesis set to 'greater', only the comparisons between the tissue with the highest abundance and the rest of tissues are included in the table. For both humans and ferrets, the lowest discrimination is observed between heart and skeletal muscle (green highlighting); the clustering process described in the methods binned these two categories together. Cluster 7 in **Figure 1c** were those transcripts that did not exhibit tissue specific expression.

Supplementary Table 10. Ferret-human sequence divergence analysis of transcription factors in tissue abundant clusters.

Tissue cluster	Number of transcription factors in the cluster	Alternative hypothesis for Mann-Whitney test		
		two sided	greater	less
Brain	30	7.40E-06	1	3.70E-06
Testis	27	0.66	0.67	0.33
Skeletal muscle, Heart	42	6.80E-06	1	3.40E-06
Lung	100	1.10E-07	1	5.60E-08
Liver	17	0.044	0.98	0.022
Kidney	43	1.20E-07	1	5.80E-08

Transcription factors (TF) in each cluster shown here were identified based on Gene Ontology annotation of human genes (sequence-specific DNA binding transcription factor activity; GO:0003700). The ferret-human divergence distances (PAM) of TFs in each cluster were compared to that of the rest of the genes in the ferret genome using Mann-Whitney two-sample test. The table shows the p-values with different alternative hypotheses to indicate if the ferret-human divergence distances of TFs in a cluster were smaller (or larger) than that of the rest of the genes in the ferret genome, indicating smaller (or larger) divergence of these TFs from human orthologs.

Supplementary Table 13. Summary of the expanded ferret annotations derived from RNA-seq data

Source	Type	Gene	Transcript
Ensembl	Protein-coding gene	19,910	20,062
	ncRNA	3,631	3,631
	Pseudogene	270	270
	Total	23,811	23,963
mRNA-seq	Novel isoforms	14,609	181,683
	Intergenic	13,379	25,975
Total RNA-seq	Intergenic	27,291	27,696
Total		64,481	259,317

Note: All transcripts predicted from mRNA-seq were spliced (having at least two exons). All predicted novel isoforms from mRNA-seq data happened to be from Ensembl annotated protein-coding genes (highlighted in Bold).

Supplementary Table 14. Summary of ferret microarray design

Annotation source		Number of genes covered (version 1)				Number of genes covered (version 2)			
		1 probe	2 probes	3 probes	Total	1 probe	2 probes	3 probes	Total
Ensembl	Protein coding	6,323	13,581	0	19,904	0	5,117	14,787	19,904
	Other	1,279	2,399	0	3,678	0	912	2,766	3,678
Intergenic, mRNA-seq		4,641	8,727	0	13,368	0	3,621	9,747	13,368
Intergenic, Total RNA-seq		0	0	0	0	0	7,038	20,250	27,288

Note: At the time of manufacturing, version 1 probe set was fitted into the Agilent 8x60k format and version 2 probe set was fitted into Agilent 4x180k format. In both cases, probes were randomly selected to fill additionally available spots, so probes were not uniformly replicated.

Supplementary Table 16. Functional enrichment analysis of lung-specific and tracheal-specific DE genes that show statistically significant differences between 1918 and CA04.

Signaling Pathways	p-value
<i>Lung-specific – 152 DE genes</i>	
Human Embryonic Stem Cell Pluripotency	9.17E-05
Extrinsic Prothrombin Activation Pathway	1.35E-03
Glioblastoma Multiforme Signaling	1.58E-03
Wnt/ β -catenin Signaling	2.75E-03
Regulation of the Epithelial-Mesenchymal Transition Pathway	3.37E-03
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	3.67E-03
Intrinsic Prothrombin Activation Pathway	4.13E-03
Coagulation System	6.41E-03
Airway Inflammation in Asthma	1.37E-02
Role of Cytokines in Mediating Communication between Immune Cells	1.53E-02
<i>Tracheal-specific – 2592 DE genes</i>	
Apoptosis Signaling	4.28E-07
NGF Signaling	1.61E-06
Ceramide Signaling	3.76E-06
Sertoli Cell-Sertoli Cell Junction Signaling	8.87E-06
IL-15 Signaling	3.87E-05
Huntington's Disease Signaling	5.23E-05
FAK Signaling	5.35E-05
Phospholipase C Signaling	7.45E-05
Acute Myeloid Leukemia Signaling	8.70E-05
Integrin Signaling	9.78E-05

As noted in Supplementary Figure 18, the lung-specific cluster consisted of 152 DE genes and the tracheal-specific cluster consisted of 2592 DE genes. Functional analysis was performed using Ingenuity Pathway Analysis.

III. Supplementary Note

1. Ferret genome assembly

Three adult sable female ferrets (*Mustela putorius furo*) (421 days old) obtained from Marshall Farms (via John Engelhardt, University of Iowa) were sacrificed and specimens sent to the Broad Institute for heterozygosity testing. The individual ID#1420 was selected for sequencing due to its low heterozygosity. The ferret DNA was sequenced to 162X total coverage by Illumina sequencing technology, and was comprised of 66X coverage using 180 bp fragment libraries, 48X coverage using 3-14kb sheared jumping libraries, and 48X coverage using ShARC jumping libraries. The reads were assembled into MusPutFur1.0 (Accession # AEYP00000000.1) using ALLPATHS-LG¹. The *M. putorius furo* genome has previously been reported to have a karyotype of 40 chromosomes². The draft assembly is 2.41 Gb in size and is composed of 2.28 Gb of sequence plus gaps between contigs. The relatively low % of gaps in this assembly speaks to its general high quality. It is composed of 117,483 contigs in 7,783 scaffolds. The longest contig is 772 kb in length, and the longest scaffold is 50.5 Mb in length. The ferret genome assembly has a contig N50 size of 44.8 kb and a scaffold N50 size of 9.3 Mb. The ambiguous base rate is 1/1929 bp, the ambiguity event rate is 1/2959, the SNP event rate is 1/6542 and the indel event rate is 1/5404. Of the 2.28 Gb in sequence in the MusPutFur1.0 assembly, 2.03 Gb of it (89%) is single copy sequence – not segmentally duplicated or a transposable element.

In addition, we sequenced a male adult sable ferret from Marshall Farms (via John Engelhardt, University of Iowa) in a separate technology development project. It was sequenced and assembled as above, with the exception that 1X of a 40 kb insert Fosill library³ was sequenced in place of the ShARC jumping library. The male ferret assembly (MusPutFurMale1.0, Accession# GCA_000239315.1) is 2.40 Gb in size and is composed of 2.33 Gb of sequence plus gaps between contigs. It has a contig N50 size of 66.7 kb and a scaffold N50 size of 9.6 Mb. 98.8% of the male ferret genome assembly aligns to the female

ferret assembly and 99.6% of the female ferret genome assembly aligns to the male assembly, the difference due mainly to the presence of the Y chromosome in the male assembly.

When repeatmasked and aligned by LastZ⁴ to the repeat masked partially finished dog assembly CanFam3.1, 85% of the non-repeat bases of MusPutFur1.0 align, which speaks to the completeness of the ferret assembly, seeing as the two species in question are separated by 43 million years. Conversely, 90% of the non-repeat bases of CanFam3.1 align to MusPutFur1.0 by LastZ.

2. Expanded annotation for differential expression analysis

To increase the coverage of our ferret transcriptome profiling analysis, we augmented Ensembl ferret genome annotation by leveraging collected RNA-seq data (Methods). As shown in Supplementary Table 13, we identified ~40,000 transcribed genomic regions which were located in between Ensembl annotated genes, and ~180,000 putative novel isoforms for 14,609 Ensembl annotated protein coding genes. These un-annotated intergenic regions encoded transcripts much longer than annotated small non-coding RNAs (Supplementary Figure 7A). Compared to Ensembl annotated protein-coding transcripts, the six-frame stop to stop translation of these intergenic transcripts showed the maximum lengths of all potential open reading frames (ORFs) are much shorter (Supplementary Figure 7B), indicating these intergenic transcripts are less likely to encode regular proteins. Instead, the results suggest these intergenic transcripts are more likely to be long non-coding RNAs, or encode small short peptides. For the novel isoforms predicted for Ensembl annotated protein-coding genes, the sense strand three-frame start to stop translation showed similar ORF lengths as Ensembl annotated isoforms (Supplementary Figure 7B), indicating many of these novel isoforms likely cover full length proteins. It's worthy of noting that Ensembl ferret annotation has about one transcript per gene at this moment, the inclusion of these novel isoforms expanded the coverage of ferret transcripts by ~10 times (Supplementary Table 13).

Next, we investigated if we captured non-polyadenylated transcripts through Total RNA-seq analysis. We reasoned that for the same gene in the same sample if Total RNA-seq analysis collected much more short reads than mRNA-seq analysis, then that gene likely transcribed non-polyadenylated transcripts, since by polyT priming mRNA-seq analysis selected against non-polyadenylated transcripts. For this purpose, we performed both Total RNA-seq and mRNA-seq analysis of 21 ferret lung samples. To facilitate the comparison, the raw gene read counts were first preprocessed as follows: *i*) any gene with less than 50 raw read counts in all 42 RNA-seq measurements were removed to ensure genes to be compared were robustly detected at least once in the samples used here, and *ii*) all gene raw read counts were scaled by the total read counts of remaining genes in each RNA-seq analysis for each sample. Next, for each gene we counted the number of samples (out of 21 samples in total) in which the scaled read count from Total RNA-seq analysis was much larger (1.5 fold or more) than that from the corresponding mRNA-seq analysis.

As expected, for the majority of annotated protein-coding genes, there were no samples for which Total RNA-seq collected more read counts than mRNA-seq (Supplementary Figure 8A), since most protein-coding transcripts are known to be polyadenylated. Pseudogenes behaved similarly as protein-coding genes (Supplementary Figure 8B), in agreement with Ensembl's definition that pseudogenes are similar to known proteins but contain a frameshift and/or stop codon(s) which disrupts the ORF. But for Ensembl annotated non-coding RNAs (snRNA, snoRNA, rRNA, miRNA, miscRNA, retrotransposed), the majority of them had many more reads in Total RNA-seq analysis than mRNA-seq analysis across all 21 samples (Supplementary Figure 8C), consistent with the current knowledge that these ncRNAs are usually non-polyadenylated. Together these results showed that this strategy was able to differentiate polyadenylated vs. non-polyadenylated transcripts.

Then we looked at intergenic transcripts. As shown in Supplementary Figure 8D, intergenic transcripts derived from Total RNA-seq data behaved similarly as Ensembl annotated

ncRNAs, i.e., the majority of them had much more reads in Total RNA-seq analysis than mRNA-seq analysis across all 21 samples, indicating these Total RNA-seq derived intergenic transcripts did indeed tend to encode non-polyadenylated transcripts. As shown in Supplementary Figure 8E, intergenic transcripts from mRNA-seq data had a bimodal distribution, suggesting that many of those regions mainly transcribed polyadenylated transcripts, but some of regions were likely to be bimorphic, i.e. the same locus transcribed both polyadenylated and non-polyadenylated transcripts. In summary, the custom annotation significantly expanded Ensembl ferret annotation by: 1) additional protein-coding isoforms, 2) non-coding RNA enriched long intergenic transcripts, and 3) non-polyadenylated long intergenic transcripts.

3. Influenza virus detection by titer, RT-PCR and immunohistochemical staining

To assess infection efficiency and compare the replication levels of CA04 and 1918 viruses in the ferret upper and lower respiratory tract, we quantified virus titers from tracheal and lung tissue homogenates, respectively (Supplementary Table 12). Similar titers were observed in all 1918 or CA04 virus-infected tracheal specimens on day 1 and 3 p.i. (mean titers for each day were not significantly different by Student's T-test, p -value > 0.1 for both time points) (Supplementary Table 12), indicating similar infection efficiency and no differences in the amount of upper respiratory tract replication between the two viruses. In the lung, the 1918 virus was recovered from all tissues on days 1 and 3 p.i., with titers on day 3 similar to that observed in other studies⁵. In contrast, no infectious virus was recovered from any of the CA04-infected lung tissues on day 1 p.i., and only 1 of 3 tissues harbored measureable virus on day 3. These data are consistent with results from previous studies with CA04 and other closely related 2009 pandemic influenza viruses^{6, 7}, in which infectious virus recovery from lungs of infected ferrets was variable. On day 8 p.i., no infectious viruses were isolated from any of the tracheal or lung samples, likely reflecting at least partial resolution of infection under both conditions. The positive infection of ferret trachea and lung tissues by 1918 and CA04 viruses were also confirmed by immunohistochemical staining of influenza antigens (Supplementary

Figures 5 and 6), clearly supporting the notion that both the 1918 and CA04 viruses spread into the pulmonary region and targets respiratory epithelium of the terminal bronchioles and the epithelium of the bronchial glands. In addition to the positive immunoreactivity for influenza antigens, viral mRNA was detected by RT-PCR from most samples where infectious virus was not recovered, with greater abundance for ferret #275 (Supplementary Table 12). Taken together, these data indicate successful infection of ferrets with 1918 and CA04 viruses.

4. Ferret microarray performance assessment

To broaden the utility of the sequenced ferret genome, we designed two versions of ferret-specific microarray for the public use (Methods, Supplementary Table 14). For both versions of microarray we submitted all Ensembl annotated genes (coding and non-coding) for probe design to maximize the coverage of Ensembl annotation. Two versions of microarray mainly differed in terms of the coverage of intergenic regions. Since the standard protocol uses polyT priming for cDNA synthesis, on version 1 microarray we added intergenic regions derived from mRNA-seq analysis, which also used polyT priming. On version 2 microarray, we included intergenic regions from both mRNA-seq and Total RNA-seq analysis in order to cover both polyadenylated and non-polyadenylated transcripts, then we experimented the cDNA synthesis with random priming.

To evaluate the performance of our custom ferret microarrays, we randomly selected 4 samples from the influenza study: lung and trachea samples from one 1918 infected ferret, and lung and trachea samples from one mock infected (PBS) control animal. For each biological sample and each version of microarray, we performed two technical replicates: aliquots of labeled cDNA from the same biological sample were hybridized to two separate microarrays.

Direct comparisons of technical replicates showed that the ferret microarray data were highly reproducible (Supplementary Figures 9 and 10). For all four pairs of technical replicates, the correlations were 0.99 or above. Across both versions of the microarray, over 90% (92.6-

96.3%) of genes had intensity differences between technical replicates that were less than 1.5 fold. For the version 1 microarray, genes with intensities that differed by 1.5 fold or larger between technical replicates were concentrated in the low intensity region (Supplementary Figure 9), which is expected for microarrays constructed with polyT priming for cDNA synthesis. For the version 2 microarray, genes with intensities that differed by 1.5 fold or larger between technical replicates tended to be more spread out (Supplementary Figure 10). It is unclear if this is due to the stochastic nature of random priming for cDNA synthesis used here, as random priming is not the standard manufacture protocol.

To evaluate if the microarrays performed well for the purpose of differential expression analysis, we calculated the infection/mock ratios based on microarray measurements, and compared them to the corresponding infection/mock ratios generated by RNA-seq analysis of the same pairs of samples. These comparisons show that infection/mock ratios from microarray and RNA-seq measurements were consistently highly positively correlated in all comparisons for both versions of the microarrays (Supplementary Figures 11-16). The overall correlations ranged from 0.71 to 0.77 (Panel B in Supplementary Figures 11-16), which is comparable to the range of 0.73 to 0.8 we observed before in comparisons of mouse microarray and RNA-seq data⁸. The correlations for the genes differentially expressed during infections were even higher, ranging from 0.78 to 0.85, across all comparisons and microarrays (Panel D in Supplementary Figures 11-16). Both versions of microarray had similar agreements with RNA-seq overall, though version 2 microarray was marginally better in terms of correlations. In general, version 2 microarray offers a broader coverage of ferret transcriptome with a potential higher per array cost, and the random priming protocol is not manufacturer's standard practice. Version 1 microarray offers comparable coverage as commonly used microarrays, with standard experimental protocol. Both versions of microarray are publicly available from the commercial vender (Agilent design IDs: 048471 and 048472). Together, these results show that the first

versions of ferret-specific microarrays are valuable genomic tools for comprehensively profiling ferret transcriptome.

References

1. Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**, 1513-1518 (2011).
2. Cavagna, P., Menotti, A. & Stanyon, R. Genomic homology of the domestic ferret with cats and humans. *Mamm Genome* **11**, 866-870 (2000).
3. Williams, L.J. et al. Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res* **22**, 2241-2249 (2012).
4. Harris, R.S. Improved pairwise alignment of genomic DNA. *Ph.D. Thesis, The Pennsylvania State University* (2007).
5. Watanabe, T. et al. Viral RNA polymerase complex promotes optimal growth of 1918 virus in the lower respiratory tract of ferrets. *Proc Natl Acad Sci U S A* **106**, 588-592 (2009).
6. Rowe, T. et al. Modeling host responses in ferrets during A/California/07/2009 influenza infection. *Virology* **401**, 257-265 (2010).
7. Itoh, Y. et al. In vitro and in vivo characterization of new swine-origin H1N1 influenza viruses. *Nature* **460**, 1021-1025 (2009).
8. Peng, X. et al. Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. *MBio* **1**, e00206-00210 (2010).